An open text authored by the Actuarial Community

Loss Data Analytics
Second Edition

Contents

Preface

Date: 19 September 2024

Book Description

Loss Data Analytics is an interactive, online, freely available text.

- The online version contains many interactive objects (quizzes, computer demonstrations, interactive graphs, video, and the like) to promote *deeper learning*.
- A subset of the book is available for *offline reading* in pdf and EPUB formats.
- The online text will be available in multiple languages to promote access to a *worldwide audience*.

What will success look like?

The online text will be freely available to a worldwide audience. The online version will contain many interactive objects (quizzes, computer demonstrations, interactive graphs, video, and the like) to promote deeper learning. Moreover, a subset of the book will be available in pdf format for low-cost printing. The online text will be available in multiple languages to promote access to a worldwide audience.

How will the text be used?

This book will be useful in actuarial curricula worldwide. It will cover the loss data learning objectives of the major actuarial organizations. Thus, it will be suitable for classroom use at universities as well as for use by independent learners seeking to pass professional actuarial examinations. Moreover, the text will also be useful for the continuing professional development of actuaries and other professionals in insurance and related financial risk management industries.

Why is this good for the profession?

An online text is a type of open educational resource (OER). One important benefit of an OER is that it equalizes access to knowledge, thus permitting a broader community to learn about the actuarial profession. Moreover, it

has the capacity to engage viewers through active learning that deepens the learning process, producing analysts more capable of solid actuarial work.

Why is this good for students and teachers and others involved in the learning process? Cost is often cited as an important factor for students and teachers in textbook selection (see a recent post on the $400 textbook). Students will also appreciate the ability to "carry the book around" on their mobile devices.

Why loss data analytics?

The intent is that this type of resource will eventually permeate throughout the actuarial curriculum. Given the dramatic changes in the way that actuaries treat data, loss data seems like a natural place to start. The idea behind the name *loss data analytics* is to integrate classical loss data models from applied probability with modern analytic tools. In particular, we recognize that big data (including social media and usage based insurance) are here to stay and that high speed computation is readily available.

Project Goal

The project goal is to have the actuarial community author our textbooks in a collaborative fashion. To get involved, please visit our Open Actuarial Textbooks Project Site.

Acknowledgements

Edward Frees acknowledges the John and Anne Oros Distinguished Chair for Inspired Learning in Business which provided seed money to support the project. Frees and his Wisconsin colleagues also acknowledge a Society of Actuaries Center of Excellence Grant that provided funding to support work in dependence modeling and health initiatives. Wisconsin also provided an education innovation grant that provided partial support for the many students who have worked on this project.

We acknowledge the Society of Actuaries for permission to use problems from their examinations.

We thank Rob Hyndman, Monash University, for allowing us to use his excellent style files to produce the online version of the book.

We thank Yihui Xie and his colleagues at Rstudio for the R bookdown package that allows us to produce this book.

We also wish to acknowledge the support and sponsorship of the Interna-

tional Association of Black Actuaries in our joint efforts to provide actuarial educational content to all.

IABA
INTERNATIONAL ASSOCIATION OF BLACK ACTUARIES

Contributors

The project goal is to have the actuarial community author our textbooks in a collaborative fashion. The following contributors have taken a leadership role in developing *Loss Data Analytics*.

Zeinab Amin

- **Zeinab Amin** is a Professor at the Department of Mathematics and Actuarial Science and Associate Provost for Assessment and Accreditation at the American University in Cairo (AUC). Amin holds a PhD in Statistics and is an Associate of the Society of Actuaries. Amin is the recipient of the 2016 Excellence in Academic Service Award and the 2009 Excellence in Teaching Award from AUC. Amin has designed and taught a variety of statistics and actuarial science courses. Amin's current area of research includes quantitative risk assessment, reliability assessment, general statistical modelling, and Bayesian statistics.
- **Katrien Antonio**, KU Leuven

Jean-François Bégin

- **Jean-François Bégin** is an Assistant Professor in the Department of Statistics and Actuarial Science at Simon Fraser University in British Columbia, Canada. Bégin holds a PhD in Financial Engineering from HEC Montréal, Canada, and is a Fellow of the Society of Actuaries and of the Canadian Institute of Actuaries. His current research interests include financial modelling, financial econometrics, Bayesian statistics, filtering methods, credit risk, option pricing, and pension economics. Bégin has designed and taught a variety of actuarial finance and actuarial communication courses.
- **Jan Beirlant**, KU Leuven

Arthur Charpentier

- **Arthur Charpentier** is a professor in the Department of Mathematics at the Université du Québec á Montréal. Prior to that, he worked at a large general insurance company in Hong Kong, China, and the French Federation of Insurers in Paris, France. He received a MS on mathematical economics at Université Paris Dauphine and a MS in actuarial science at ENSAE (National School of Statistics) in Paris, and a PhD degree from KU Leuven, Belgium. His research interests include econometrics, applied probability and actuarial science. He has published several books (the most recent one on *Computational Actuarial Science with R*, CRC) and papers on a variety of topics. He is a Fellow of the French Institute of Actuaries, and was in charge of the 'Data Science for Actuaries' program from 2015 to 2018.

Curtis Gary Dean

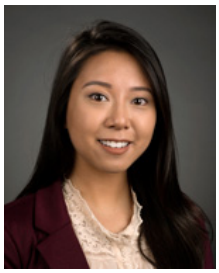
- **Curtis Gary Dean** is the Lincoln Financial Distinguished Professor of Actuarial Science at Ball State University. He is a Fellow of the Casualty Actuarial Society and a CFA charterholder. He has extensive practical experience as an actuary at American States Insurance, SAFECO, and Travelers. He has served the CAS and actuarial profession as chair of the Examination Committee, first editor-in-chief for *Variance: Advancing the Science of Risk*, and as a member of the Board of Directors and the Executive Council. He contributed a chapter to *Predictive Modeling Applications in Actuarial Science* published by Cambridge University Press.

Edward (Jed) Frees

- **Edward (Jed) Frees** is an emeritus professor, formerly the Hickman-Larson Chair of Actuarial Science at the University of Wisconsin-Madison. He is a Fellow of both the Society of Actuaries and the American Statistical Association. He has published extensively (a four-time winner of the Halmstad and Prize for best paper published in the actuarial literature) and has written three books. He also is a co-editor of the two-volume series *Predictive Modeling Applications in Actuarial Science* published by Cambridge University Press.

Guojun Gan

- **Guojun Gan** is an associate professor in the Department of Mathematics at the University of Connecticut, where he has been since August 2014. Prior to that, he worked at a large life insurance company in Toronto, Canada for six years. He received a BS degree from Jilin University, Changchun, China, in 2001 and MS and PhD degrees from York University, Toronto, Canada, in 2003 and 2007, respectively. His research interests include data mining and actuarial science. He has published several books and papers on a variety of topics, including data clustering, variable annuity, mathematical finance, applied statistics, and VBA programming.



Lisa Gao

- **Lisa Gao** is a PhD candidate in the Risk and Insurance department at the University of Wisconsin-Madison. She holds a BMath in Actuarial Science and Statistics from the University of Waterloo and is an Associate of the Society of Actuaries.
- **José Garrido**, Concordia University

Lei (Larry) Hua

- **Lei (Larry) Hua** is an Associate Professor of Actuarial Science at Northern

Illinois University. He earned a PhD degree in Statistics from the University of British Columbia. He is an Associate of the Society of Actuaries. His research work focuses on multivariate dependence modeling for non-Gaussian phenomena and innovative applications for financial and insurance industries.

Noriszura Ismail

- **Noriszura Ismail** is a Professor and Head of Actuarial Science Program, Universiti Kebangsaan Malaysia (UKM). She specializes in Risk Modelling and Applied Statistics. She obtained her BSc and MSc (Actuarial Science) in 1991 and 1993 from University of Iowa, and her PhD (Statistics) in 2007 from UKM. She also passed several papers from Society of Actuaries in 1994. She has received several research grants from Ministry of Higher Education Malaysia (MOHE) and UKM, totaling about MYR1.8 million. She has successfully supervised and co-supervised several PhD students (13 completed and 11 on-going). She currently has about 180 publications, consisting of 88 journals and 95 proceedings.

Joseph H.T. Kim

- **Joseph H.T. Kim**, Ph.D., FSA, CERA, is Associate Professor of Applied Statistics at Yonsei University, Seoul, Korea. He holds a Ph.D. degree in Actuarial Science from the University of Waterloo, at which he taught as Assistant Professor. He also worked in the life insurance industry. He has published papers in *Insurance Mathematics and Economics*, *Journal of Risk and Insurance*, *Journal of Banking and Finance*, *ASTIN Bulletin*, and *North American Actuarial Journal*, among others.

Nii-Armah Okine

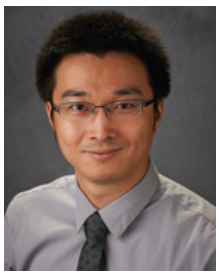
- **Nii-Armah Okine** is an assistant professor at the Mathematical Sciences Department at Appalachian State University. He holds a Ph.D. in Business (Actuarial Science) from the University of Wisconsin - Madison and obtained his master's degree in Actuarial science from Illinois State University. His research interest includes micro-level reserving, joint longitudinal-survival modeling, dependence modeling, micro-insurance, and machine learning.

Rajesh (Raj) Sahasrabuddhe

- **Rajesh (Raj) Sahasrabuddhe** is a Partner and Philadelphia Office Leader with Oliver Wyman Actuarial Consulting. Raj is a Fellow of the Casualty Actuarial Society (CAS), an Associate of the Canadian Institute of Actuaries, and a Member of the American Academy of Actuaries. Raj has been an active volunteer with CAS Admissions committees throughout his career, including a term as Chairperson of the Syllabus Committee from 2010 to 2013. He currently serves on the MAS-II Examination Committee. He has authored or co-authored papers that have appeared on syllabi for both the CAS and Society of Actuaries.

Emine Selin Sarıdaş

- **Emine Selin Sarıdaş** is a doctoral candidate in the Statistics department of Mimar Sinan University. She holds a bachelor degree in Actuarial Science with a minor in Economics and a master degree in Actuarial Science from Hacettepe University. Her research interest includes dependence modeling, regression, loss models and life contingencies.

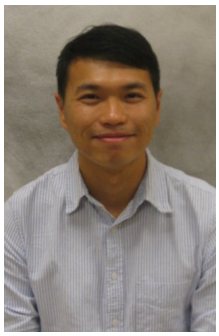


Peng Shi

- **Peng Shi** is an associate professor in the Risk and Insurance Department at the Wisconsin School of Business. He is also the Charles & Laura Albright Professor in Business and Finance. Professor Shi is an Associate of the Casualty Actuarial Society (ACAS) and a Fellow of the Society of Actuaries (FSA). He received a Ph.D. in actuarial science from the University of Wisconsin-Madison. His research interests are problems at the intersection of insurance and statistics. He has won several research awards, including the Charles A. Hachemeister Prize, the Ronald Bornhuetter Loss Reserve Prize, and the American Risk and Insurance Association Prize.

Nariankadu D. Shyamalkumar (Shyamal)

- **Nariankadu D. Shyamalkumar (Shyamal)** is an associate professor in the Department of Statistics and Actuarial Science at The University of Iowa. He is an Associate of the Society of Actuaries, and has volunteered in various elected and non-elected roles within the SoA. Having a broad theoretical interest as well as interest in computing, he has published in prominent actuarial, computer science, probability theory, and statistical journals. Moreover, he has worked in the financial industry, and since then served as an independent consultant to the insurance industry. He has experience educating actuaries in both Mexico and the US, serving in the roles of directing an undergraduate program, and as a graduate adviser for both masters and doctoral students.



Jianxi Su

- **Jianxi Su** is an Assistant Professor at the Department of Statistics at Purdue University. He is the Associate Director of Purdue's Actuarial Science. Prior to joining Purdue in 2016, he completed the PhD at York University (2012-2015). He obtained the Fellow of the Society of Actuaries (FSA) in 2017. His research expertise are in dependence modelling, risk management, and pricing. During the PhD candidature, Jianxi also worked as a research associate at the Model Validation and ORSA Implementation team of Sun Life Financial (Toronto office).

Chong It Tan

- **Chong It Tan** is a senior lecturer at Macquarie University in Australia, where he has served as the undergraduate actuarial program director since 2018. He obtained his PhD in 2015 from Nanyang Technological University in Singapore. He is a fully qualified actuary, holding the credentials from both the US Society of Actuaries and Australian Actuaries Institute. His major research interests are mortality modelling, longevity risk management and bonus-malus systems.

Tim Verdonck

- **Tim Verdonck** is associate professor at the University of Antwerp. He has a degree in Mathematics and a PhD in Science: Mathematics, obtained at the University of Antwerp. During his PhD he successfully took the Master in Insurance and the Master in Financial and Actuarial Engineering, both at KU Leuven. His research focuses on the adaptation and application of robust statistical methods for insurance and finance data.

Krupa Viswanathan

- **Krupa Viswanathan** is an Associate Professor in the Risk, Insurance and Healthcare Management Department in the Fox School of Business, Temple University. She is an Associate of the Society of Actuaries. She teaches courses in Actuarial Science and Risk Management at the undergraduate and graduate levels. Her research interests include corporate governance of insurance companies, capital management, and sentiment analysis. She received her Ph.D. from The Wharton School of the University of Pennsylvania.

Reviewers

Our goal is to have the actuarial community author our textbooks in a collaborative fashion. Part of the writing process involves many reviewers who generously donated their time to help make this book better. They are:

- Yair Babab
- David Back, Liberty Mutual
- Chunsheng Ban, Ohio State University
- Vytaras Brazauskas, University of Wisconsin - Milwaukee
- Yvonne Chueh, Central Washington University
- Chun Yong Chew, Universiti Tunku Abdul Rahman (UTAR)
- Benjamin Côté, Université Laval
- Eren Dodd, University of Southampton
- Gordon Enderle, University of Wisconsin - Madison
- Rob Erhardt, Wake Forest University
- Runhun Feng, University of Illinois
- Brian Hartman, Brigham Young University
- Liang (Jason) Hong, University of Texas at Dallas
- Fei Huang, Australian National University
- Hirokazu (Iwahiro) Iwasawa

- Himchan Jeong, University of Connecticut
- Min Ji, Towson University
- Paul Herbert Johnson, University of Wisconsin - Madison
- Dalia Khalil, Cairo University
- Samuel Kolins, Lebonan Valley College
- Andrew Kwon-Nakamura, Zurich North America
- Ambrose Lo, University of Iowa
- Mélina Mailhot, Concordia University
- Mark Maxwell, University of Texas at Austin
- Tatjana Miljkovic, Miami University
- Bell Ouelega, American University in Cairo
- Zhiyu (Frank) Quan, University of Connecticut
- Jiandong Ren, Western University
- Margie Rosenberg, University of Wisconsin - Madison
- Rajesh V. Sahasrabuddhe, Oliver Wyman
- Sherly Paola Alfonso Sanchez, Universidad Nacional de Colombia
- Ranee Thiagarajah, Illinois State University
- Ping Wang, Saint Johns University
- Chengguo Weng, University of Waterloo
- Toby White, Drake University
- Michelle Xia, Northern Illinois University
- Di (Cindy) Xu, University of Nebraska - Lincoln
- Lina Xu, Columbia University
- Lu Yang, University of Amsterdam
- Chun Yong
- Jorge Yslas, University of Copenhagen
- Jeffrey Zheng, Temple University
- Hongjuan Zhou, Arizona State University

Other Collaborators

- Alyaa Nuval Binti Othman, Aisha Nuval Binti Othman, and Khairina (Rina) Binti Ibraham were three of many students at the Univeristy of Wiscinson-Madison that helped with the text over the years.
- Maggie Lee, Macquarie University, and Anh Vu (then at University of New South Wales) contributed the end of the section quizzes.
- Jeffrey Zheng, Temple University, Lu Yang (University of Amsterdam), and Paul Johnson, University of Wisconsin-Madison, led the work on the glossary.

Version Number

- This is **Version 2.0**, October 2024. Edited by Hélène Cossette, Edward (Jed) Frees, Brian Hartman, and Tim Higgins.
- Version 1.1, August 2020. Edited by Edward (Jed) Frees and Paul Johnson.
- Version 1.0, January 2020, was edited by Edward (Jed) Frees.

You can also access pdf and epub (current and older) versions of the text in our Offline versions of the text.

For our Readers

We hope that you find this book worthwhile and even enjoyable. For your convenience, at our Github Landing site (https://openacttexts.github.io/), you will find links to the book that you can (freely) download for offline reading, including a pdf version (for Adobe Acrobat) and an EPUB version suitable for mobile devices. Data for running our examples are available at the same site.

In developing this book, we are emphasizing the online version that has lots of great features such as a glossary, code and solutions to examples that you can be revealed interactively. For example, you will find that the statistical code is hidden and can only be seen by clicking on terms such as

We hide the code because we don't want to insist that you use the `R` statistical software (although we like it). Still, we encourage you to try some statistical code as you read the book – we have opted to make it easy to learn `R` as you go. We have set up a separate R Code for Loss Data Analytics site to explain more of the details of the code.

Like any book, we have a set of notations and conventions. It will probably save you time if you regularly visit our Appendix Chapter **??** to get used to ours.

Freely available, interactive textbooks represent a new venture in actuarial education and we need your input. Although a lot of effort has gone into the development, we expect hiccoughs. Please let your instructor know about opportunities for improvement, write us through our project site, or contact chapter contributors directly with suggested improvements.

1

Appendix. Data Resources

This appendix section describes the datasets used in this book and others that you may wish to explore.

For each set of data, we provide download buttons so that you can easily access the data in standard .csv (comma separated value) format. This allows you replicate and experiment with the methods developed in the book as well as sharpen your understanding through exercises.

We provide the source of each dataset. We also recommend, for deeper understanding, that you occasionally refer to these original sources to further develop your appreciation of the data underpinning the analytics developed in this book.

1.1 Wisconsin Property Fund

Description: The Wisconsin Local Government Property Insurance Fund (LGPIF) is an insurance pool administered by the Wisconsin Office of the Insurance Commissioner. The LGPIF was established to provide property insurance for local government entities that include counties, cities, towns, villages, school districts, and library boards. The fund insures local government property such as government buildings, schools, libraries, and motor vehicles. It covers all property losses except those resulting from flood, earthquake, wear and tear, extremes in temperature, mold, war, nuclear reactions, and embezzlement or theft by an employee.

The data are available using this download button: Download the Wisconsin Property Fund Data

TABLE 1.1: **Variables in the Wisconsin Property Fund Dataset**

Variable	Description
PolicyNum	Policy number
Year	Contract year
Premium	Premium
Deduct	Deductible
BCcov	Coverage for building and contents
Freq	Number of claims during the year (frequency)
Fire5	Binary variable to indicate the fire class is below 5
NoClaimCredit	Binary variable to indicate no claims in the past two years
EntityType	Categorical variable that is one of six types: 1=Village, 2=City,3=County, 4=Misc, 5=School, or Town)
AlarmCredit	Categorical variable that is one of four types: (0, 5, 10, or 15) for automatic smoke alarms in main rooms
BCClaim	Builing and contents claims

TABLE 1.2: **Wisconsin Property Fund First Five Rows**

PolicyNum	Year	Premium	Deduct	BCcov	Freq	Fire5	NoClaimCredit	EntityType	AlarmCredit	BCClaim
120002	2006	9313	1000	22714456	0	1	0	3	1	0
120002	2007	8767	1000	25046646	0	1	0	3	1	0
120002	2008	7090	1000	20851525	0	1	1	3	1	0
120002	2009	8522	1000	21852696	0	1	1	3	1	0
120002	2010	7994	1000	23511493	1	1	1	3	1	6839

TABLE 1.3: **Wisconsin Property Fund Last Five Rows**

PolicyNum	Year	Premium	Deduct	BCcov	Freq	Fire5	NoClaimCredit	EntityType	AlarmCredit	BCClaim
180787	2010	199	500	285000	0	1	1	4	1	0
180788	2010	58344	100000	416739800	1	1	0	4	1	168304
180789	2010	295	500	500988	1	1	0	4	1	1034
180790	2010	2077	1000	3580665	0	1	0	4	4	0
180791	2010	81	500	118800	0	1	0	4	1	0

1.2 ANU Corporate Travel Data

Universities purchase corporate travel policies to cover employees and students traveling on official university business for a wide variety of accidents and incidents while away from the campus or primary workplace. This broad coverage includes medical care and evacuation, loss of personal property, extraction for political and weather related reasons, and more. See Frees and Butt (2022) for more information about this coverage.

There are 2107 observations in this dataset. The variable names are described in Table 1.4 and the first and last five observations are in Table 1.6.

Data are available using this button: Download Corporate Travel Claims Data.

TABLE 1.4: **Variables in the Corporate Travel Dataset**

Variable	Description
UW Year	Underwriting Year
Loss Date	Date that the loss occurred
Reported Date	Date that the loss was reported
Last Trans Date	Last date in which there was a transaction regarding the loss
Paid Loss	Cumulative amount paid on the loss
Outstanding Reserve	Estimate of the loss amount yet to be paid
Incurred Loss	Sum of the amount paid and the estimate of future payments
Status	An indicator as to whether the claim has been deemed settled (closed) or not settled (open)

TABLE 1.5: **Corporate Travel Data First Five Rows**

UW.Year	Loss.Date	Reported.Date	Last.Trans.Date	Paid.Loss	Outstanding.Reserve	Incurred.Loss	Status
2021	19/12/2021	20/12/2021	24/12/2021	10000	0	10000	Closed
2021	9/4/2022	29/04/2022	30/05/2022	423	0	423	Closed
2021	2/5/2022	4/5/2022		0	500	500	Open
2021	5/5/2022	17/05/2022		0	562	562	Open
2021	30/04/2022	27/05/2022	10/6/2022	1500	0	1500	Closed

TABLE 1.6: **Corporate Travel Data Last Five Rows**

UW.Year	Loss.Date	Reported.Date	Last.Trans.Date	Paid.Loss	Outstanding.Reserve	Incurred.Loss	Status
2006	1/11/2006	19/06/2007		0	0	0	Closed
2006	24/06/2007	26/06/2007	8/1/2008	6278	0	6278	Closed
2006	4/7/2007	6/7/2007	11/9/2007	114	0	114	Closed
2006	20/05/2007	26/06/2007	14/07/2007	136	0	136	Closed
2006	15/02/2007	27/06/2007	14/07/2007	1208	0	1208	Closed

Source: Frees, Edward and Butt, Adam (2022). "ANU Corporate Travel Insurance Claims 2022". Australian National University Data Commons. DOI https://doi.org/10.25911/vrdw-9f32.

1.3 ANU Group Personal Accident Data

Group personal accident insurance offers financial protection in case of injury or death resulting from an incident that occurs on the job. Like workers' compensation, group personal accident offers insurance coverage and liability insurance protection against accidental death or injury. Unlike workers' compensation, group personal accident covers students and ANU's voluntary workers. See Frees and Butt (2022) for more information about this coverage.

There are 148 observations in this dataset. The variable names are described in Table 1.7 and the first and last five observations are in Table 1.9.

Data are available using this button: Download Group Personal Accident Claims Data.

TABLE 1.7: **Variables in the Group Personal Accident Dataset**

Variable	Description
UW Year	Underwriting Year
Loss Date	Date that the loss occurred
Last Trans Date	Last date in which there was a transaction regarding the loss.
Paid Loss	Cumulative amount paid on the loss
Outstanding Reserve	Estimate of the loss amount yet to be paid
Incurred Loss	Sum of the amount paid and the estimate of future payments
Status	An indicator as to whether the claim has been deemed settled (closed) or not settled (open)

TABLE 1.8: **Group Personal Accident Data First Five Rows**

UW.Year	Loss.Date	Last.Trans.Date	Paid.Loss	Outstanding.Reserve	Incurred.Loss	Status
2021	6/12/2021	3/6/2022	805	0	805	Closed
2021	15/11/2021		0	0	0	Closed
2021	15/11/2021		0	0	0	Closed
2021	22/03/2022	4/5/2022	396	0	396	Closed
2021	11/4/2022	2/8/2022	740	360	1100	Open

TABLE 1.9: **Group Personal Accident Data Last Five Rows**

UW.Year	Loss.Date	Last.Trans.Date	Paid.Loss	Outstanding.Reserve	Incurred.Loss	Status
2010	6/3/2011	26/07/2011	776	0	776	Closed
2010	22/07/2011	23/01/2012	4625	0	4625	Closed
2010	5/6/2011	30/01/2012	1504	0	1504	Closed
2007	11/1/2008	23/02/2008	0	0	0	Closed
2007	29/08/2008		0	0	0	Closed

Source: Frees, Edward and Butt, Adam (2022). "ANU Group Personal Accident Claims 2022". Australian National University Data Commons. https://doi.org/10.25911/jcfx-zj56.

1.4 ANU Motor Vehicle Data

This policy covers ANU's vehicles including cars, vans, utilities, and motorcycles. See Frees and Butt (2022) for more information about this coverage.

There are 318 observations in this dataset. The variable names are described in Table 1.10 and the first and last five observations are in Table 1.12.

Data are available using this button: Download Motor Vehicle Claims Data.

TABLE 1.10: **Variables in the Motor Vehicle Dataset**

Variable	Description
Policy Term Start Date	Start date of the contract year in which the loss occurred
Loss Date	Date that the loss occurred
Reported Date	Date that the loss was reported
Motor Fault	Party responsible for the loss
Driver Age	Age of the driver
Vehicle Description	Type of vehicle
Loss Postcode	Postal code where the loss occurred
Excess	The deductible applied to the loss
Motor Net Paid	Amount paid to the insured (ANU)
Outstanding Estimate	Estimate of the loss amount yet to be paid
Motor Net Incurred	Sum of the amount paid and the estimate of future payments
Third Party Identified	Indicates whether a responsible third party could be identified
Third Party Insured	Indicates whether a responsible third party was insured

TABLE 1.11: **Motor Vehicle Data First Five Rows**

Policy.Term.Start.Date	Loss.Date	Reported.Date	Motor.Fault	Driver.Age	Vehicle.Description	Loss.Postcode
1/11/2011	6/6/2012	4/10/2012	THIRD PARTY RESPONSIBLE	NA	FORD TRANSIT VAN	2600
1/11/2011	16/08/2012	14/11/2013	INSURED RESPONSIBLE	39	TOYOTA HIACE	2612
1/11/2011	4/9/2012	17/01/2013	INSURED RESPONSIBLE	52	HYUNDAI IX35	2600
1/11/2011	21/09/2012	28/09/2012	THIRD PARTY RESPONSIBLE	59	HOLDEN COMMODORE	2518
1/11/2011	22/09/2012	12/10/2012	INSURED RESPONSIBLE	NA	SUBARU FORESTER	2612

Excess	Motor.Net.Paid	Outstanding.Estimate	Motor.Net.Incurred	Third.Party.Identified	Third.Party.Insured
1000	385	0	385	IDENTIFIED	
1000	901	0	901		
1000	1226	0	1226		
NA	1672	0	1672	IDENTIFIED	NOT INSURED
1000	3419	0	3419		INSURED

Source: Frees, Edward and Butt, Adam (2022). "ANU Motor Vehicle Claims 2022". Australian National University Data Commons. DOI https://doi.org/10.25911/g7e4-9e46.

TABLE 1.12: **Motor Vehicle Data Last Five Rows**

Policy.Term.Start.Date	Loss.Date	Reported.Date	Motor.Fault	Driver.Age	Vehicle.Description	Loss.Postcode
1/11/2021	4/4/2022	5/4/2022	INSURED RESPONSIBLE	66	VOLKSWAGEN TIGUAN	2604
11/1/2021	11/4/2022	9/5/2022	INSURED RESPONSIBLE	27	TOYOTA HILUX	2540
1/11/2021	11/4/2022	9/5/2022	INSURED RESPONSIBLE	27	TOYOTA HILUX	2540
11/1/2021	15/04/2022	11/7/2022	INSURED RESPONSIBLE	21	TOYOTA HILVX	2601
1/11/2021	18/07/2022	18/07/2022	NO-ONE RESPONSIBLE	NA	TOYOTA HILUX	2601

Excess	Motor.Net.Paid	Outstanding.Estimate	Motor.Net.Incurred	Third.Party.Identified	Third.Party.Insured
0	2373	1056	3429		
0	210	25000	25210		
0	0	31927	31927		
0	0	2750	2750		
0	0	299	299		

1.5 Spanish Personal Insurance Data

This dataset consists of 10,000 insurance private customers of a real portfolio of insurance policy holders in Spain with a motor insurance and a homeowners insurance contract for policy year 2014. The data contain information on each customer, policies and yearly claims by type of contract.

The data are available using this download button: Download the Spanish Personal Insurance Data

The description of the data appears in Table 1.13.

TABLE 1.13: Variable and Description of Spanish Personal Insurance Data

Variable	Description
gender	1 for male and 0 for female
Age_client	the age of the customer in years
year	Policy year. Equals 5 corresponding to 2014.
age_of_car_M	the number of years since the vehicle was bought by the customer
Car_power_M	the power of the vehicle
Car_2ndDriver_M	1 if the customer has informed the insurance company that a second occasional driver uses the vehicle, and 0 otherwise
num_policiesC	the total number of policies held by the same customer in the insurance company
metro_code	1 for urban or metropolitan and 0 for rural
Policy_PaymentMethodA	1 for annual payment and 0 for monthly payment in the motor policy
Policy_PaymentMethodH	1 for annual payment and 0 for monthly payment in the homeowners policy
Insuredcapital_content_re	the value of content in homeowners insurance
Insuredcapital_continent_re	the value of building in homeowners insurance
appartment	1 if the homeowners insurance correspond to an apartment and 0 otherwise
Client_Seniority	the number of years that the customer has been in the company
Retention	1 if the policy is renewed and 0 otherwise
NClaims1	the number of claims in the motor insurance policy for the corresponding year
NClaims2	the number of claims in the homeowners insurance policy for the corresponding year
Claims1	the sum of claims cost in the motor insurance policy for the corresponding year
Claims2	the sum of claims cost in the homeowners insurance policy for the corresponding year
Types	1 when neither an auto nor a home claim, it is equal to 2 when the customer has an auto but not a home claim, it is equal to 3 when the customer does not have not an auto but a home claim and it is equal to 4 when both an auto and a home claim.
PolID	Policy Identification Number

All monetary units are expressed in Euros. In motor insurance, only claims at fault are considered.

These data were drawn from a larger database of 40,284 insurance private customers. These customers are tracked from 2010 to 2014. Some customers do not renew their policies, so that they do not stay in the sample for five years. For the smaller data, only the 2014 policy year was used and from this, a random sample of 10,000 customers was drawn.

TABLE 1.14: **Spanish Personal Insurance Data First Five Rows**

gender	Age.client	year	age.of.car.M	Car.power.M	Car.2ndDriver.M	um.policiesC
1	47	5	12	163	0	0
1	52	5	13	80	0	1
0	66	5	7	97	0	1
1	70	5	17	95	0	1
1	67	5	13	110	0	1

metro.code	Policy.PaymentMethodA	Policy.PaymentMethodH	Insuredcapital.content.re	Insuredcapital.continent.re	appartment
0	1	1	10	12	1
0	1	1	10	11	0
1	1	1	9	11	1
0	1	1	10	11	1
0	1	1	11	12	0

Client.Seniority	Retention	NClaims1	NClaims2	Claims1	Claims2	Types	PolID
7	1	0	0	0	0	1	12476
18	1	0	0	0	0	1	29232
15	1	0	0	0	0	1	23770
16	1	0	1	0	58	3	8228
6	1	0	0	0	0	1	37088

TABLE 1.15: **Spanish Personal Insurance Data Last Five Rows**

gender	Age.client	year	age.of.car.M	Car.power.M	Car.2ndDriver.M	um.policiesC
1	66	5	8	143	0	1
1	55	5	18	125	1	1
0	41	5	10	190	0	1
1	50	5	5	140	0	1
1	55	5	12	90	0	1

metro.code	Policy.PaymentMethodA	Policy.PaymentMethodH	Insuredcapital.content.re	Insuredcapital.continent.re	appartment
0	1	1	10	11	1
0	1	1	11	11	1
0	1	1	9	12	1
0	1	1	10	12	0
1	1	1	11	13	0

Client.Seniority	Retention	NClaims1	NClaims2	Claims1	Claims2	Types	PolID
20	1	0	0	0	0	1	2967
15	1	0	0	0	0	1	9387
6	1	0	0	0	0	1	36519
8	1	0	0	0	0	1	33276
6	1	0	0	0	0	1	25370

See Frees et al. (2021) for more information about this dataset. The larger database contains 122935 rows and is freely available at:

Source: Guillen, Montserrat; Bolancé, Catalina; Frees, Edward W.; Valdez, Emiliano A. (2021), "Insurance data for homeowners and motor insurance customers monitored over five years", Mendeley Data, V1, DOI https://doi.org/10.17632/vfchtm5y7j.1

1.6 'R' Package CASdatasets

The `R` package `CASdatasets` provides a convenient way to access many well-known insurance datasets. This package was originally created to support the book *Computational Actuarial Science with R*, edited by Arthur Charpentier, Charpentier (2014).

To install the package, here is a bit of `R` code:

```
install.packages("CASdatasets", repos = "http://cas.uqam.ca/pub/", type = "source")
library(CASdatasets)
`?`(CASdatasets)
`?`(sgautonb  # See the documentation of the Singapore Auto Data
)
`?`(lossalae  # See the documentation of the Loss and Expense Data
)
```

Note that this package assumes that you have already installed a few other packages, including *xts*, *sp*, and *zoo*.

To illustrate,

- in Chapter **??** we use the Singapore data (referred to as `sgautonb` in the package) and
- in Chapter **??** we use the loss and expense data (referred to as `lossalae` in the package).

1.7 Other Data Sources

There exists man other (non-actarial) data sources. First, data can be obtained from university-based researchers who collect primary data. Second, data can be obtained from organizations that are set up for the purpose of releasing secondary data for the general research community. Third, data can be obtained from national and regional statistical institutes that collect data. Finally, companies have corporate data that can be obtained for research purposes.

While it might be difficult to obtain data to address a specific research problem or answer a business question, it is relatively easy to obtain data to test a model or an algorithm for data analysis. In the modern era, readers can obtain

datasets from the Internet. The following is a list of some websites to obtain real-world data:

- **UCI Machine Learning Repository.** This website (url: http://archive.ics.uci.edu/ml/index.php) maintains more than 400 datasets that can be used to test machine learning algorithms.
- **Kaggle.** The Kaggle website (url: https://www.kaggle.com/) include real-world datasets used for data science competitions. Readers can download data from Kaggle by registering an account.
- **DrivenData.** DrivenData aims at bringing cutting-edge practices in data science to solve some of the world's biggest social challenges. In its website (url: https://www.drivendata.org/), readers can participate in data science competitions and download datasets.
- **Analytics Vidhya.** This website (url: https://datahack.analyticsvidhya.com/contest/all/) allows you to participate and download datasets from practice problems and hackathon problems.
- **KDD Cup.** KDD Cup is the annual Data Mining and Knowledge Discovery competition organized by the ACM Special Interest Group on Knowledge Discovery and Data Mining. This website (url: http://www.kdd.org/kdd-cup) contains the datasets used in past KDD Cup competitions since 1997.
- **U.S. Government's open data.** This website (url: https://www.data.gov/) contains about 200,000 datasets covering a wide range of areas including climate, education, energy, and finance.
- **AWS Public Datasets.** In this website (url: https://aws.amazon.com/datasets/), Amazon provides a centralized repository of public datasets, including some huge datasets.

Bibliography

Charpentier, Arthur (2014). *Computational actuarial science with R*. CRC press.

Frees, Edward W, Catalina Bolancé, Montserrat Guillen, and Emiliano A Valdez (2021). "Dependence modeling of multivariate longitudinal hybrid insurance data with dropout," *Expert Systems with Applications*, Vol. 185, p. 115552.

Frees, Edward W and Adam Butt (2022). "ANU Insurable Risks," URL: https://doi.org/10.25911/0SE7-N746.